LEARNING THE SPIRAL SHARING NETWORK WITH MINIMUM SALIENT REGION REGRESSION FOR SALIENCY DETECTION

Zukai Chen[*1,2], *Xin Tan*[1,2], *Hengliang Zhu*[1,2], *Shouhong Ding*[4], *Lizhuang Ma*[1,2,3] *and Haichuan Song*[*2]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
[3]School of Computer Science and Software Engineering, East China Normal Univeristy
[4]Tencent Youtu Lab
banmaczk@sjtu.edu.cn, hcsong@sei.ecnu.edu.cn

ABSTRACT

With the development of convolutional neural networks (CNNs), saliency detection methods have made a big progress in recent years. However, the previous methods sometimes mistakenly highlight the non-salient region, especially in complex backgrounds. To solve this problem, a two-stage method for saliency detection is proposed in this paper. In the first stage, a network is used to regress the minimum salient region (RMSR) containing all salient objects. Then in the second stage, in order to fuse the multi-level features, the spiral sharing network (SSN) is proposed for pixel-level detection on the result of RMSR. Experimental results on four public datasets show that our model is effective over the state-of-the-art approaches.

Index Terms— saliency detection, salient region, spiral sharing network

1. INTRODUCTION

Saliency detection refers to the extraction of significant areas in images by simulating human visual characteristics. As a classic computer vision task, it is usually used as a pretreatment stage for many applications, such as object detection[1], object tracking[2], image classification[3] and semantic segmentation[4]. Although there are many researches[5, 6, 7, 8, 9, 10, 11] about saliency detection, it is still challenging to distinguish the salient objects with complex backgrounds.

Many existing saliency detection methods directly use the raw image to model saliency. There are noticeable problems in these methods. 1) It is difficult to tell the saliency with objects in the same category in the image. As shown in the first row of Fig. 1, there are many people in the image while only the center girl is the salient one. Two previous saliency detection methods (UCF[10] and C2S[11]) mistakenly identify that the rest people are also salient. 2) In addition, when the background is similar to the salient objects in color distribution, it is not easy to distinguish salient objects. As illustrated in the second row of Fig. 1, a yellow dog is sleeping on a yellow blanket, but those two methods highlight the blanket by mistake.

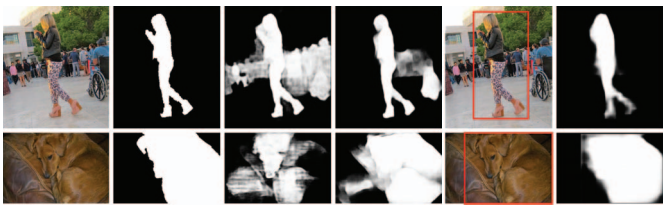


Fig. 1. Examples of the weakness in previous methods and the importance of an ideal salient region. From left to right: image, ground truth mask, UCF [10], C2S[11], our first-stage (RMRN) results and our final saliency maps.

For saliency detection, location is as important as expansion. While in many end-to-end CNN-based methods[8, 11, 12], they take the raw image for pixel-level detection to accomplish these two tasks at the same time, but it is difficult to locate the salient region especially in complex backgrounds. Hence, a two-stage model is utilized in our work. In the first stage, a network is used to regress the minimum salient region (RMSR) which can remove the background to the maximum extent and also preserve all salient objects, which is also called salient region detection. As Fig. 1 (e) shows, RMSR locates the salient region even in the complicated environment. In this case, we can focus on the expansion in the second stage.

It is knwon that multi-level information has also different contributions for saliency detection. Low-level information contains some texture features which help to preserve

*Corresponding author. Thanks to the Science and Technology Commission of Shanghai Municipality Program (No. 18D1205903) and National Natural Science Foundation of China (No. 61872242, 61472245) for funding.

the boundary and shape of salient object, while it cannot help salient objects stand out from the background. To explore the semantic properties of salient objects, high-level information should also be taken into account. As above, both high-level and low-level information are important for saliency detection. Therefore, the spiral sharing network (SSN) with a spiral structure connected by skip-layer block is proposed for to combine different level information.

Based on the above motivations, a new saliency detection model is designed to extract the salient objects from complex backgrounds. As shown in Fig. 1(f), even if the background of the image is complex, our two-stage model still can obtain a better result. Our work has three contributions:

- A two-stage saliency detection method is proposed to help salient objects stand out from complex backgrounds. Experimental results of several datasets show that our method is effective.
- In the first stage, we utilize a network to regress the minimum salient region (RMSR) containing all salient objects so that these objects can be separated from the background.
- In the second stage, to obtain more information, the spiral sharing network (SSN) is proposed for pixel-level saliency detection that can utilize high-level and low-level features.

2. OUR APPROACH

In this paper, a two-stage method is proposed for saliency detection in complex backgrounds. As Fig. 2 shows, in the first stage, in order to locate the salient region, we take a network to regress the minimum salient region (MSR). In the second stage, we feed the MSR to the network. To utilize multi-level features, a new network named spiral sharing network (SSN) is designed for the pixel-level saliency detection. From this two-stage model, we get the final saliency map.

2.1. Regression of Minimum Salient Region (RMSR)

In this work, salient region which contains all salient objects, is located by two coordinates: the upper-left coordinate (W_{min}, H_{min}) and the lower-right coordinate (W_{max}, H_{max}), There is a definition for the region as:

$$f(x_i, y_j) = 0 \qquad (x_i, y_j) \notin MSR, \tag{1}$$

where (x_i, y_j) represents a pixel in the image. And if the pixel is not in MSR, it is not salient and equals to 0. Also, it requires that at least one salient pixel equaling to 1 exists on each boundary of MSR. A deep convolutional neural network (i.e. general VGG16[13]) is applied to regress this region, and the ground truth is the bounding box which can be easily obtained by the existing saliency datasets. And its loss

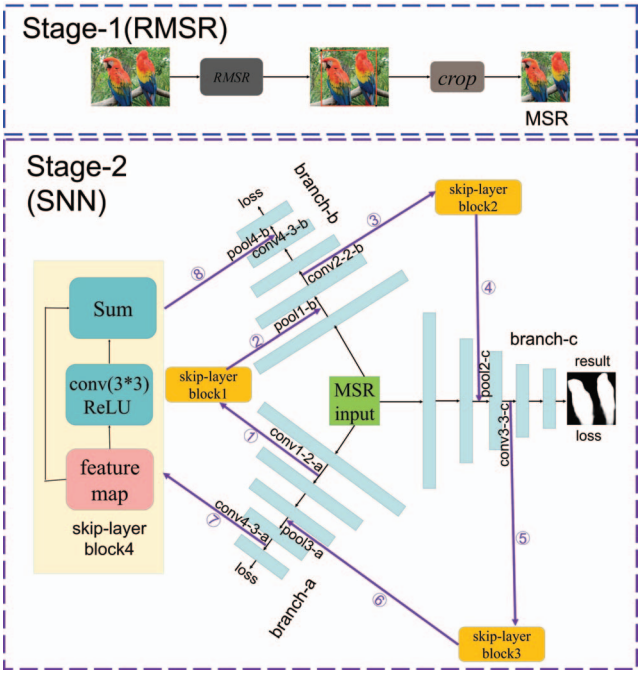


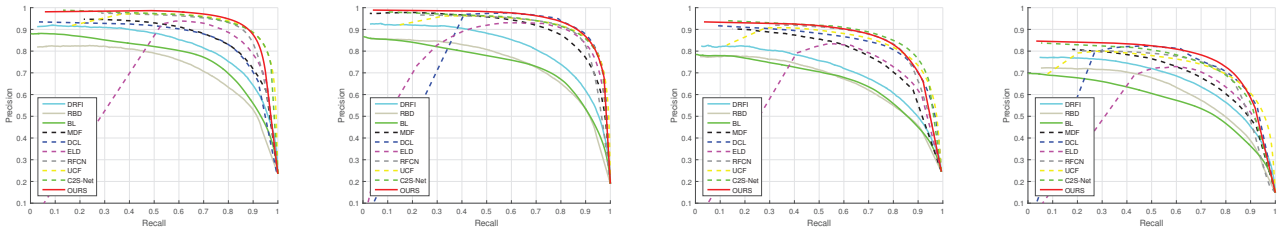
Fig. 2. The structure of our two-stage model for saliency detection. In the Stage-1, RMSR network is utilized to regress the minimum salient region (MSR). In the Stage-2, the three-branch networks embedded with skip-layer block on each branch is used to transfer and share multi-level information.

funchtion is Euclidean loss. With the RMSR, we can separate the salients objects from the complex background. As the stage-1 of Fig. 2 shows, the result of the RMSR is also robust even there are multiple salient objects in the image.

2.2. Spiral Sharing Network (SSN)

As shown in the stage-2 of Fig. 2, the input of SSN is the cropped image according to the output of RMSR. The structure of SSN consists of three branches. In order to share more information between different networks, a spiral structure, which consists of four skip-layer blocks, is embedded into the network. With this structure, our model can share multiple features from different branches and layers. It starts from conv1-2-a of branch-a, through skip-layer block1, passing the information from branch-a to branch-b. Similarly, the information of branch-b is passed to branch-c through skip-layer block2. With these four blocks, a spiral structure is formed among the three branches so that the low-level features in the foregoing branch are combined with the high-level features in the next branch. CrossEntropy loss is used to calculate the gap between saliency map and ground truth. In our case, we also take the VGG16[13] as the branch network. Although it is a very simple network in many tasks, it still performs well with our spiral structure which is described in Sec. 3.

Skip layer block, referring to the idea of ResNet[16], com-



(a) ECSSD (b) HKU-IS (c) PASCAL-S (d) DUT-OMRON

Fig. 3. Compare our proposed algorithm with other state-of-the-art methods on ECSSD, HKU-IS, PASCAL-S and DUT-OMRON datasets via PR curves.

Table 1. Comparison of our proposed algorithm with other state-of-the-art methods via F-measure value(the higher the better) and MAE (the lower the better) based on ECSSD, HKU-IS, PASCAL-S and DUT-OMRON datasets. The best results are shown in red and sub-optimal results are shown in blue.

Method	ECSSD		HKU-IS		PASCAL-S		DUT-OMRON	
	maxFm	MAE	maxFm	MAE	maxFm	MAE	maxFm	MAE
DRFI [5]	0.786	0.164	0.783	0.143	0.690	0.281	0.664	0.150
RBD [14]	0.716	0.171	0.726	0.141	0.655	0.273	0.630	0.141
BL [15]	0.755	0.217	0.723	0.206	0.659	0.318	0.580	0.240
MDF [6]	0.831	0.105	0.861	0.129	0.764	0.142	0.694	0.092
DCL [9]	0.901	0.075	0.907	0.055	0.810	0.115	0.756	0.086
ELD [7]	0.868	0.079	0.881	0.063	0.771	0.121	0.705	0.091
RFCN [8]	0.898	0.095	0.888	0.080	0.832	0.118	0.738	0.095
UCF [10]	0.903	0.069	0.888	0.061	0.818	0.116	0.730	0.120
C2S [11]	0.900	0.057	0.886	0.050	0.840	0.089	0.737	0.080
OURS	0.910	0.062	0.900	0.048	0.834	0.103	0.762	0.068

bines the information between different layers. For example, as shown in the skip-layer block4 of Fig. 2, the output of this block is the sum of the feature maps of the 3*3 convolution layer with the non-linear transformation (ReLU[17]) and previous feature maps which keeps the low-level information. The input of this block is the feature maps of conv4-3-a and the output of this block passes to pool4-b which belongs to branch-b. It can be formulated as:

$$I_{p4-b} = SUM(F_{sk-b-4} + F_{c4-3-b}), \tag{2}$$

where I_{p4-b} is the input of pool1-b. F_{sk-b-4} and F_{c4-3-b} are the outputs of the skip-layer block-4 and conv4-3-b respectively.

In addition, in the training phase of SSN, since the network is a complete three-branch structure, there are three losses in the network: one of them is master loss and the others are auxiliaries. However, in the test, because the output of branch-c is the final saliency map, the layers that irrelevant to this output will be removed aiming at simplifying the testing network. In this way, the test speed can be improved.

3. EXPERIMENT

3.1. Benchmark Datasets and Evaluation Metrics

Dataset. We test our method on ECSSD[18], PASCAL-S[19], HKU-IS[20] and DUT-OMRON[21]. Many images have more than one salient object in these datasets.

Evaluation Metrics. To evaluate our method, we use three metrics: precision-recall curves (PR curve), F-measure[22], and the mean absolute error (MAE). The PR curve can be obtained by comparing the binary mask of salient map with ground truth. F-measure is the weighted average of precision and recall which is formulated as:

$$F_m = \frac{(1+m^2) \times precision \times recall}{m^2 \times precision + recall}, \tag{3}$$

where $m^2 = 0.3$ just like many salient works[7, 10, 11]. The MAE is used to measure the difference between the predicted value of the classifier and the actual result, the smaller the better. We can represent it as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |S_i - G_i|^2, \tag{4}$$

where S_i and G_i represent the result and the ground truth of one pixel which are normalized to [0,1]. The results of these metrics are described in sec. 3.3.

3.2. Implementation

The work is implemented with the publicly Caffe[23] library. Both of the two networks are train on MSRA-B[24] and MSRA10K[25] datasets. In the first stage, the image is scaled to 300×300 uniformly because of the full connected layers. Then according to the ground truth of saliency map, two coordinates (mentioned in section2.1), which can locate the MSR, is labeled as the ground truth for this stage. In the second stage, only the minimum salient region in the image is utilized as our training data. To improve the recall rate of MSR, so that the cropped image can contain all salient objects, we expand the result of RMSR 0.2 times as input for the second stage. Finally, the cropped image is padded to the size of the original image after the second stage. VGG16[13] is chosen as the pre-trained model for both two networks. The hyper-parameters used in this work contains: learning policy (step), base learning rate (le-8), step-size (5000), momentum (0.90) and weight decay (0.0008).

3.3. Performance Comparison

Our saliency detection method is compared with six CNN-based methods: MDF[6],DCL[9], ELD[7], RFCN[8], UCF[10], C2S[11] as well as with three classical methods: DRFI[5], RBD[14], BL[15]. For a fair comparison, we use the saliency results or source code provided by the author.

PR curve. As shown in Fig. 3, our method is compared with the methods mentioned above through PR curves. In all datasets, our method performs well compared to other methods at the beginning, because the salient map we get is very close to the ground truth. Even if when the recall rate is very high at the end, our prediction rate is still competitive. This shows that our algorithm can be adapted to a variety of complex scenarios.

F-measure and MAE. F-measure and MAE of methods are shown in Tabel. 1. On four datasets, our algorithm achieves top two over both F-measure and MAE. Especially our algorithm has the best MAE values on HKU-IS and DUT-OMRON datasets. As for the F-measure, our method performs best over F-measure in ECSSD and DUT-OMRON. This result shows that our method is robust to separate salient objects from complex backgrounds.

Visual comparison. As shown in Fig. 4, our saliency map is compared with those by other methods. It is obvious that our approach is able to highlight the salient object especially in complex and confused backgrounds. This is all due to the fact that we located the saliency regions in the first stage.

Running time. It takes 13 hours and 15 hours respectively to train our two networks on a single NVIDIA GTX-1080TI and a 3.6GHz Intel processor. During the testing, due to the clipping and splicing operations in the whole process, it takes 0.12s to process an image of size 400×300, which is faster than many CNN-based methods[6, 8, 9].

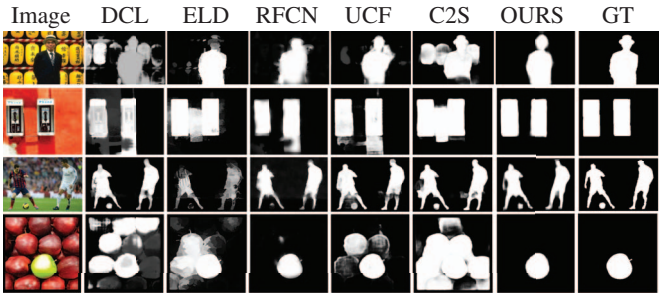


Fig. 4. Visual comparison of our proposed method with other state-of-the-art methods.

3.4. Ablation Studies

To evaluate the performance of the first stage, we use two metrics: intersection over union (IOU) and recall rate. On ECSSD and PASCAL-S datasets, IOU is 0.803 and 0.786. As for recall rate, it is 99% and 98.3% on two datasets.

We also perform a contrast experiment on ECSSD and PASCAL-S datasets to evaluate the effectiveness of our method. We take out skip-layer blocks from SSN trained with raw image as baseline model. Then we train SSN with raw image too. Finally we use the cropped datasets dealt with the RMSR as training image for baseline and SSN. As shown in Table 2, the two-stage model performs best. Compared with baseline, our method improves the maximum F-measure by 1.5% and 2.5% and decrease the MAE by 18% and 13% over on ECSSD and PASCAL-S datasets, respectively.

Table 2. Comparison of F-measure and MAE on ECSSD and PASCAL-S datasets to evaluate the performance of SSN and RMSR.

Method	ECSSD		PASCAL-S	
	maxFm	MAE	maxFm	MAE
Baseline	0.896	0.076	0.812	0.118
SSN	0.906	0.065	0.825	0.107
baseline+RMSR	0.905	0.066	0.823	0.109
SSN+RMSR	0.910	0.062	0.832	0.103

4. CONCLUSION

In this paper, we propose a two-stage method of saliency detection to get a result with clean background as well as highlighting salienct objects. The first stage is utilized to locate salient region so that we can use a limited region instead of the raw image for pixel-level saliency detection. In the second stage, a spiral sharing network is designed to share the features between different layers and multiple branches, and through these abundant information, our final saliency map is very close to ground truth. Experiment on several datasets show that our method achieves a state-of-the-art performance.

5. REFERENCES

[1] Vidhya Navalpakkam and Laurent Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," 2006, vol. 2, pp. 2049–2056.

[2] Vijay Mahadevan, Nuno Vasconcelos, et al., "Biologically inspired object tracking using center-surround saliency mechanisms.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 541–554, 2013.

[3] Ruobing Wu, Yizhou Yu, and Wenping Wang, "Scale: Supervised and cascaded laplacian eigenmaps for visual object recognition based on nearest neighbors," in *CVPR*, 2013, pp. 867–874.

[4] Michael Donoser, Martin Urschler, Martin Hirzer, and Horst Bischof, "Saliency driven total variation segmentation," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 817–824.

[5] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li, "Salient object detection: A discriminative regional feature integration approach," in *CVPR*, 2013, pp. 2083–2090.

[6] Guanbin Li and Yizhou Yu, "Visual saliency based on multiscale deep features," in *CVPR*, 2015, pp. 5455–5463.

[7] Gayoung Lee, Yu Wing Tai, and Junmo Kim, "Deep saliency with encoded low level distance map and high level features," in *CVPR*, 2016, pp. 660–668.

[8] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Ruan Xiang, "Saliency detection with recurrent fully convolutional networks," in *European Conference on Computer Vision*, 2016, pp. 825–841.

[9] Guanbin Li and Yizhou Yu, "Deep contrast learning for salient object detection," in *CVPR*, 2016, pp. 478–487.

[10] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin, "Learning uncertain convolutional features for accurate saliency detection," 2017, pp. 212–221.

[11] Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Dinggang Shen, "Contour knowledge transfer for salient object detection," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[12] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr, "Deeply supervised salient object detection with short connections," in *CVPR*, 2017, pp. 3203–3212.

[13] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[14] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun, "Saliency optimization from robust background detection," in *CVPR*, 2014, pp. 2814–2821.

[15] Na Tong, Huchuan Lu, Ruan Xiang, and Ming Hsuan Yang, "Salient object detection via bootstrap learning," in *CVPR*, 2015, pp. 1884–1892.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[18] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia, "Hierarchical saliency detection," in *CVPR*, 2013, pp. 1155–1162.

[19] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille, "The secrets of salient object segmentation," in *CVPR*, 2014, pp. 280–287.

[20] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang, "Saliency detection by multi-context deep learning," in *CVPR*, 2015, pp. 1265–1274.

[21] Chuan Yang, Lihe Zhang, Huchuan Lu, Ruan Xiang, and Ming Hsuan Yang, "Saliency detection via graph-based manifold ranking," in *CVPR*, 2013, pp. 3166–3173.

[22] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk, "Frequency-tuned salient region detection," in *CVPR*, 2009, pp. 1597–1604.

[23] Jia, Yangqing, Shelhamer, Evan, Donahue, Jeff, Karayev, Sergey, Long, and Jonathan, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Multimedia*, 2014, pp. 675–678.

[24] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung Yeung Shum, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, 2011.

[25] Ming Ming Cheng, Guo Xin Zhang, N. J Mitra, and Xiaolei Huang, "Global contrast based salient region detection," in *CVPR*, 2011, pp. 409–416.